DOCUMENT RESUME

ED 076 605 TM 002 585

AUTHOR	Davis, Richard W.; Loadman, William E.
TITLE	The Matrix Test Analysis Program: A Measurement Heuristic.
PUB DATE	Feb 73
NOTE	13p.; Paper prepared for National Council on Measurement in Education (New Orleans, Louisiana, February 25 - March 1, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Computer Programs; *Item Analysis; *Item Sampling; *Matrices; Measurement Techniques; Performance; *Psychometrics; Statistical Analysis; Technical Reports; *Test Results

ABSTRACT

A subject by item matrix of test responses is shown to be a useful heuristic in criterion referenced and norm referenced test analysis, and in the teaching of measurement. The pattern of responses within the matrix provides indications of item interactions, weak deceptors, and conventional test statistics. The strong visual analogy between the matrix and test parameters makes the matrix a useful teaching aid and analytical tool. (Author)

FORM 510

PRINTED IN U.S.A.

ED 076605

TM 002 585

10.16

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

The Matrix Test Analysis Program

A Measurement Heuristic

Richard W. Davis

Indiana University

William E. Loadman

Nisonger Center

Ohio State University

Paper Prepared for NCME Meeting

New Orleans, La.

February, 1973

INTRODUCTION

Today the psychometrician has available to him many sophisticated and informative item analysis tools. Computer programs for test analysis, such as Veldman's TESTSTAT program or Baker's Wisconsin Item Analysis Program (FORTAP), are in such general use that test authors commonly regard them as an integral part of test writing. Yet for the unsophisticated test author, the teacher in the classroom, or even the student in introductory courses in testing and measurement, test and item analysis are mysterious and largely meaningless considerations. The reason for this may be the general mathemaphobia exhibited by many educators, but whatever its source, a problem does exist. Actually the difficulty is not limited to test users. Faced with increasingly large tests which produce incredible amounts of data, the test author at any level is seeking methods to make his task of analysis less burdenson.

Still a third aspect of the test analysis problem has been introduced by the new educational technology. There one finds increased need for tests which relate on one hand to the sophisticated theoretical models in measurement, and on the other to the practical problems of supporting ongoing instruction. One example of a serious attempt to serve both the practical and theoretical aspects of testing is found in the Roudebush and Green (1971) paper in which they present a

strategy for using a matrix of student responses derived from the "categorical and hierarchical structure" of the content domain over which the measure is constructed to represent the results of a diagnostic measure in mathematics for the primary grades. Some have applied the term blueprint to this approach. In addition to its ability to present relatively complex data meaningfully, the Roudebush and Green idea is interesting because it involves a return to "eyeballing" the data, albeit at a remarkably sophisticated level.

In the light of current, sophisticated, quantitative models of test and item performance, their matrix seems novel in the sense that in it one actually looks at the individual item response as the fundamental unit of the test analysis. Of course the test itself has been validated in a more or less conventional manner, and the matrix does not replace the validation function of test construction. Still the matrix provides at an immediate, useful level a great deal of information which would be very difficult to get at through conventional item statistics.

In the scheme Roudebush and Green put forward each hierarchy of the test is represented as a column vector in the item matrix. Each row corresponds to some level of the hierarchy. Of particular interest is the idea, not made explicit in the paper, that any one hierarchy performs

essentially like a Guttman scale since one expects an individual to respond correctly up to some maximum level (corresponding to the individual's ability) and to miss on all items higher in the hierarchy. Unfortunately, work with true Guttman scales applied to measuring cognitive abilities has been almost uniformly dissappointing. First, the requirement of the model that the items have very high discriminatory power is rarely met. The result of that is that a scale of more than just a very few items must span a relatively broad range of ability to perform as a Guttman scale. The second problem is that it is rarely possible to construct scales with acceptably high coefficients of reproducability.

The remainder of this paper will discuss a computer program designed to provide a visual display of item and subject performance on a test (or series of tests) as well as to allow for conventional item analysis procedures and item matrix sampling procedures.

Ordered Vectors

There is a key concept here which has great merit whether or not the item hierarchies perform as Guttman scales. The real significance of the hierarchy is that more information is present in an ordered response vector than in an unordered one. What this information is and how to get at it are really the key questions being raised here.

To determine what the pattern of responses in the ordered vector means it is first necessary to consider how

the ordering is performed. In the case considered above, ordering is performed according to some pre-established task and item hierarchy. What the ordered response vector of item responses shows in this situation is the extent to which the item difficulty as indicated by the proportion of correct responses corresponds to the hierarchical ordering. One might even proceed to compute the rank order correlation coefficient between the two scales to provide a quantitative index of the correspondence. Note, however, that the point of the analysis has shifted from the previous discussion. Whereas Roudebush and Green are interested in individual subject characteristics, the purpose now is to use the total sample of individual responses to learn something about the performance of the items on the test.

The significance of the hierarchy in this new context is that one would expect the items on a reliable test to reflect the hierarchical structure present in the items. In the face of the current debate about the nature of reliability in criterion measurement one might propose that the reliability of the criterion sub-test, that is items related to a single set of tasks or a single dimension of the content domain, is a function of the degree to which the pattern of correct responding to the items corresponds to the pattern of dependencies in the hierarchy. The principal distinction to be

made between this approach and Guttman scaling is that here no assumption is being made about the discriminatory power of the individual items. It is also possible to display specific sub sets of items. Suppose a particular sub-set of items is related to a given objective. It is possible to quickly ascertain the performance of a group of subjects on one or more objectives by manipulating the order of the items so that all items related to the objectives were displayed in serial order. In this fashion the program can be used to analyze criterion (or domain) referenced measures.

Matrix Sampling Model

The correlation between the ordered ranking of the sub-test and the difficulty rankings of the items would be of considerable interest in the matrix sampling model. Among the principal benefits of the matrix sampling model is the efficient measurement of group characteristics. One of those characteristics may be the performance of a group on a measure following a unit of instruction on a hierarchically structured task. In that common situation one is seeking a means of indexing the group performance as it relates to the structure of the hierarchy, since in the formative evaluation or development of an instructional unit. It is extremely important to know if the task hierarchy as designed is actually at work in the learning environment. The significance of the ordered response vector is simply that it pro-

vides the desired information in a much more meaningful manner than a conventional unordered item matrix.

Of more general interest, perhaps, is the simpler situation in which no hierarchy is presumed to be present in the responses. Here the items are ordered only by some index of difficulty such as proportion of correct responding. But instead of looking just at the vector of all responses to the measure, suppose one were to examine the response vector for each score group. Then one would expect that for each non-zero score group the pattern of responding, again the ranking of item difficulty as indexed by the proportion of correct responses for the individuals in the score group, would correlate highly with the rank ordering of the items by difficulty for all subjects. Essentially, there should be no large shifts in the pattern of responses for any score group as compared to the total sample of individuals taking the test. If this were not the case one would have evidence that either ability on the measure as a whole interacts with ability on some of the items of the measure, or guessing is a significant source of variance.

Typically what one observes in the non-parametric correlation of score group response pattern with the pattern of responding of the entire group is that high score groups exhibit little variance in their scores (low correlation), middle range scoregroups account for most of the variance

(high correlation), while low score groups have increasingly large amounts of error variance due to guessing (very low correlation).

To make the observation of the relationships discussed above easier, one can represent the data as a matrix in which the columns are formed from the test items arranged in rank order of difficulty from highest to lowest, and in which the rows are formed from the score groups arranged in rank order from the group with the highest score to the group with the lowest. If the score group of people who get no items correct is omitted and if each subject takes each item, the matrix is rectangular. In the matrix sampling case, the number of score groups will simply equal the number of items each subject takes. In either situation the value in the i, j^{th} cell of the matrix represents the proportion of correct responses by the members of the i^{th} scoregroup to the j^{th} item. (See Table 1)

(Place Table 1 about here)

Interpreting the Matrix

The ordered response matrix presents some very interesting information in a readable format. If one looks down a column of the matrix he is looking at the item trace--the probability of getting the item correct as a function of

subject ability. This is perhaps the best indicator of item discriminatory power and has the advantage that quick comparisons of discriminatory power can be made among all the items of the test. It is interesting to note that if the items of the measure are performing like a Guttman scale, one would expect the matrix to be upper triangular with ones on and above the diagonal and zeros below.

A comparison of the rows of the matrix indicates the relative performance of the score groups and may provide useful, though non-quantitative, support for a scheme of assigning grades based on the differences between a pair of score groups. In the case of guessing one would observe a marked deterioration of the pattern for the lower score groups.

In essence then what the matrix provides is a visual analogy to many quantitatively defined test statistics. And while the experienced test analyst may find the analogy of marginal utility, particularly since there is a sacrifice of precision and since the information about test performance obtained from conventional test statistics has great meaning for him, the analogy has proven to be of great general usefulness to unsophisticated test authors who have difficulty grasping the theoretical basis of item and test analysis. They are often able to deal with problems of test design such as reliability through the visual analogy of the matrix much more readily than they are able to obtain them from the

mathematical definitions of the statistics. In addition, one has the possibility of applying conventional item analysis routines to the data as well as matrix sampling procedures developed by Shoemaker (1971), Knapp (1972) and Bunda (1971).

Subject Orderings

Thus far all groupings and orderings have referred to items or item based statistics. There is no reason, however, to exclude the possibility of ordering subjects independently of the test items. Certainly the possibility might arise in which one wishes to examine the performance of a sub-group of subjects on the basis of a treatment or aptitude defined prior to the test. In such a situation the researcher would like to ask not only about the effects due to the treatment--the typical research question--but also about whether or not the single test performs in the same manner for the two subject groups. "Performing in the same manner" can be defined in terms of a large number of interrelated statistics, and the visual matrix of items and score groups provides a good, preliminary indication of which direction the quantitative analysis might best take. This is particularly true of large or complex tests such as one frequently encounters in matrix sampling methodology which may involve the responses of several hundred subjects to as many items.

With such large amounts of data the practical value of a data representation scheme like the matrix test analysis becomes significant.

A more general case arises, however, in which one wishes to examine subject performance on a measure as a funtion of some other measure--for example one might wish to examine performance on a criterion test as a function of one or another aptitudes indexed on another measure. In such a situation the item ordering in the matrix would remain the same, while individuals would be ordered in the rows according to their performance on the second measure. Clustering of subjects would produce manageable numbers of rows. The great benefit of the matrix test analysis program in this context is its ability to present many aspects of the test performance at once. For example, patterns of systematic missing which appear independently of item difficulty could indicate an aptitude interaction. In any event the general pattern of responses would provide an indication of any aptitude effect for the items under study. In its ability to provide this kind of information about the test, the matrix test analysis scheme corresponds to test analysis very much like histograms correspond to marginal statistics and plots correspond to correlation and regression analysis.

CONCLUSION

There are, then, four basic points being made here. The first is made only implicitly. But it is important, for as the amount of data being gathered and analyzed increases, the task of simply looking at it grows geometrically. Furthermore, to the unsophisticated test user the analysis of test characteristics is a difficult and time consuming task. What is needed, therefore, is a scheme which reduces the information processing burden associated with the analysis of tests--a sort of plot of the data to provide a starting point for the quantitative analysis. For the sophisticated user there are the conventional (and not always appropriate) item analysis routines.

The second point relates to ordering. Although there are problems associated with the ordering of items and subjects, problems in defining hierarchies and problems in scaling, there is clearly more information in a set of ordered items than in a set of unordered items. The test analyst should be aware of the meaning of this extra data and attempt to use it to identify patterns and relationships among the items and subjects.

Thirdly, there is merit in working with the data directly. And, just as one should always check correlations for the possibility of non-linear data, one should check tests for trends and relationships which may not appear in coefficients

of reliability or inter-item correlations. In each of these three problem areas, processing complex data, inspection of the ordered relationships, and visual inspection of the data, the matrix test analysis scheme shows considerable promise of helping the test analyst make meaningful decisions. As such it is a useful measurement heuristic.

Finally, an item analysis program, appropriate for domain referenced measurement is available and provides the user with visual quantitative displays on item and individual and/or score group performance.

Bibliography

Bunda, M.A., "An Investigation of an Extension of Item Sampling Which Yields Individual Scores" Unpublished doctoral dissertation, University of Illinois, 1971.

Knapp, Thomas, Annotated Bibliography on Item Person Sampling. Mimeo. University of Rochester, 1972.

Roudabush, Glenn and Green, Donald, "Some Reliability Problems in a Criterion-Referenced Test" AERA Paper, New York, 1971.

Shoemaker, David, Principles and Procedures of Matrix Sampling. South West Regional Laboratory· Technical Report 34, 1971.

Table 1

Score Group by Item Matrix for a simulated ten item test
Columns are items, rows are score groups, cells are the percent of correct responses by score group on item.

ITEM n = 70

SCORE GROUP	1	2	3	4	5	6	7	8	9	10	GROUP SIZE	Spearman Correlation with item bank undefined
10	100	100	100	100	100	100	100	100	100	100	2	
9	40	100	80	80	100	100	100	100	100	100	5	.67
8	25	62	62	100	75	100	100	75	100	100	8	.75
7	29	21	50	71	86	86	71	100	86	100	14	.88
6	17	25	50	33	50	66	83	83	92	100	12	.98
5	0	20	20	40	40	0	60	80	100	100	5	.84
4	0	0	20	0	50	30	40	80	80	100	10	.92
3	0	25	50	25	0	0	0	25	75	75	4	.43
2	*	*	*	*	*	*	*	*	*	*	0	--
1	*	*	*	*	*	*	*	*	*	*	0	---
ITEM SCORE	12	19	29	31	38	38	42	51	54	59	70	

WEL:wr
2/21/73